Inappropriate Comment Classification System

Hongyi Zheng, Maggie Wang

BACKGROUND

Purpose

The Inappropriate Comment Classification System aims to correctly classify different types of inappropriate comments. The labels for inappropriate comments are:

toxic, severe toxic, threats, obscene, insults, and ***identity-based hate***.

A comment may have multiple labels.

Strengths

This ADS allows platforms to filter multiple types of inappropriate comments.

* Different website may prefer to filter different kinds of inappropriate content based on their function and property.

INPUT & PREPROCESSING

Input

The dataset is composed of comments collected from Wikipedia talk page edits.

The input features are User ID and Comment Text.

User ID: Distinct label for each users who post comment on website.

Comment Text: Text content of the comment posted.

INPUT & OUTPUT

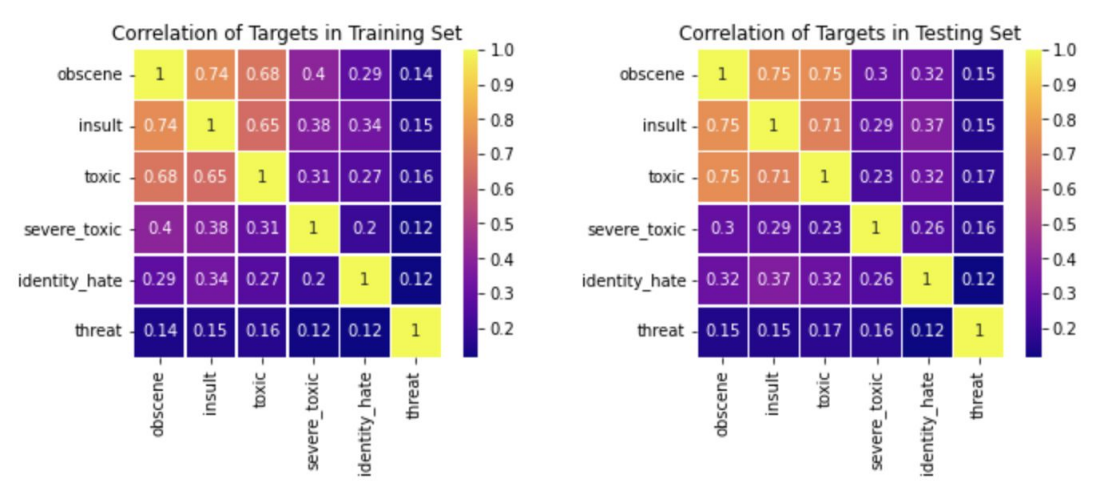
Exploratory Data Analysis

In the training set, 89.93% of the comments have negative labels for all categories, whereas some label combinations only account for 0.01% of the dataset. This discovery suggests the imbalance of the training data, which might lead to a biased model. Upsampling or downsampling might need to be implemented.

INPUT & OUTPUT

Exploratory Data Analysis

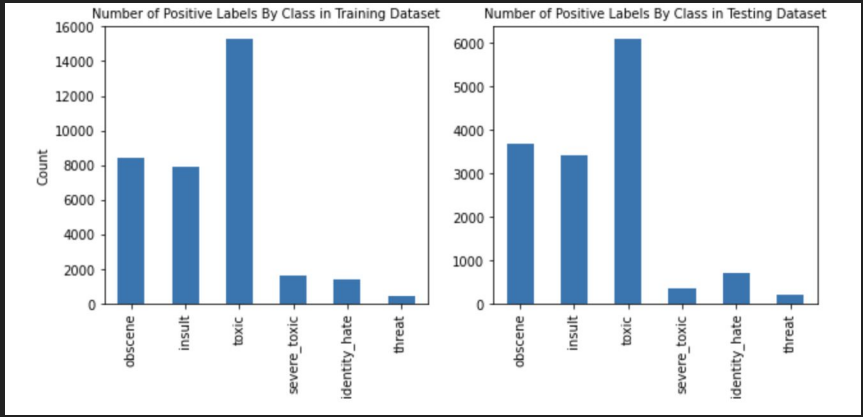
Since we only have one input feature (comment text), we instead analyze the correlation between target columns and draw two heatmaps visualizing the correlation between labels in training and test set respectively.

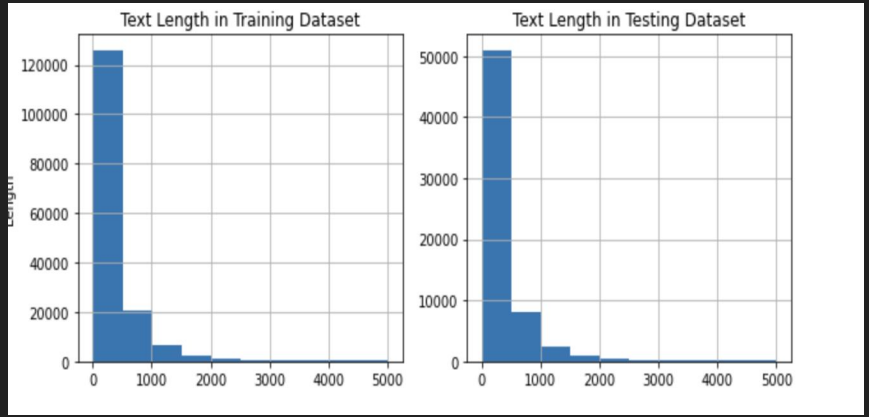


INPUT & OUTPUT

Exploratory Data Analysis

We compare the label distribution and text length in training and test set.





IMPLEMENTATION & VALIDATION

We mainly use the solution from Rhodium Beng as the guideline for our model implementation. The core of this solution are multiple logistic regression classifiers, each one of them determines whether a comment belongs to a certain type of inappropriate comments. Nevertheless, based on our needs, we also introduced multiple improvements compared with the original implementation.

This solution is simple but effective. Moreover, this low-complexity solution allow us to better interpret the underlying model with LIME to understand how it works.

We replace LogisticRegression() with LogisticRegressionCV(), adding cross-validation and grid-search of regulation parameter to the model implementation to ensure the robustness of the model.

IMPLEMENTATION & VALIDATION

Preprocessing Text

Use regex expressions and string translate() method to remove digits, punctuations and replace abbreviations.

TF-IDF Vectorizer

Term frequency-inverse document frequency. This vectorizer increases its value as the number of times a word appears increases while decreasing the value as the number of documents containing a certain word increases. It could capture words that truly contribute to the classification results by giving higher values to rare words.

SMOTE

Synthetic minority oversampling technique, one of the data augmentation methods. SMOTE can produce new synthetic example belonging to minority class in the dataset by picking up a random point between a minority class example and one of its nearest neighbors. Utilizing SMOTE solve the problem of imbalance data in training set, as mentioned in exploratory data analysis in Input & Output section.

IMPLEMENTATION & VALIDATION

Performance on Training Set

The performance of the algorithm on the training set is evaluated by accuracy metrics. Models trained with original data and upsampled data all achieve high overall accuracies, indicating the great performance of the algorithm.

	Without Upsampling	With Upsampling
Obscene	0.9817	0.9649
Insult	0.9742	0.9520
Toxic	0.9624	0.9377
Severe Toxic	0.9907	0.9823
Identity Hate	0.9930	0.9798
Threat	0.9976	0.9992

Table 1 Training Accuracy On Different Target Variables

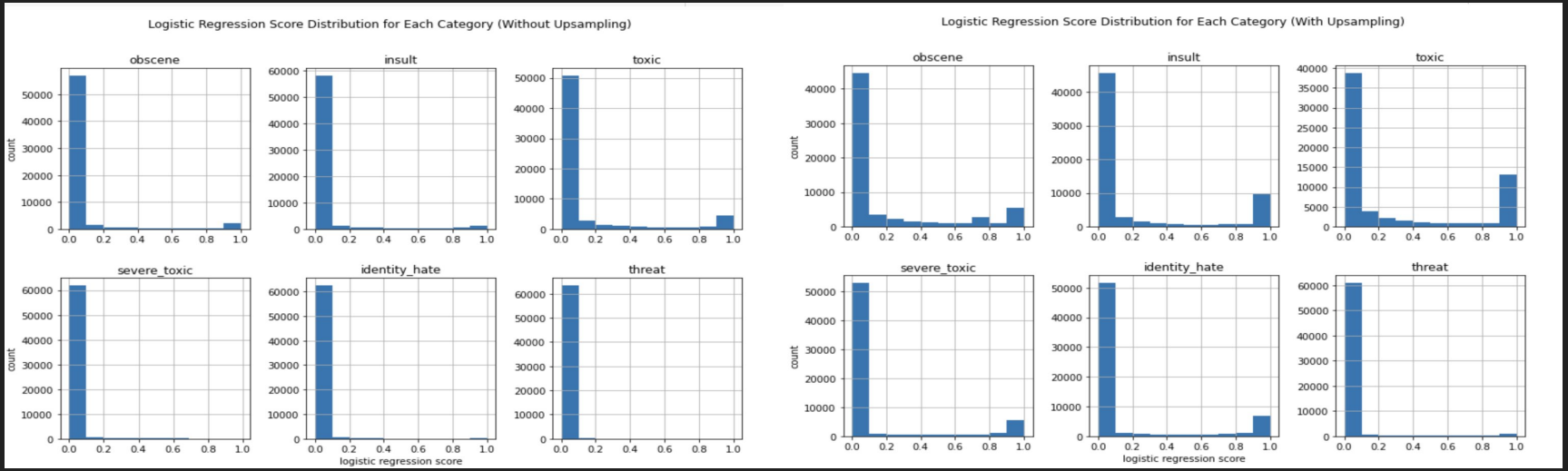
OUTPUT

The outputs of this ADS are the predicted probabilities of every comment belonging to each of the six types of inappropriate comments. If the predicted probability of a comment being a type of inappropriate comment is greater than 0.5, then we predict that comment belongs to this type of inappropriate comment. Vice versa.

For the classification purpose, we transform the probability into binary labels: mark as negative (value 0) for probability smaller than 0.5; mark as positive (value 1) for probability greater than 0.5

OUTPUT

Distribution of Logistic Regression Score



OUTCOMES

Fairness metrics

The graph on top shows the metrics of each classifier without upsampling.

The graph on bottom shows the metrics of each classifier with upsampling.

	Obscene	Insult	Toxic	Severe Toxic	Identity Hate	Threat
Accuracy	0.9654	0.9636	0.9287	0.9932	0.9910	0.9962
Recall	0.6681	0.5518	0.7534	0.2752	0.3666	0.2938
Precision	0.7140	0.7043	0.6001	0.3769	0.6779	0.4026
FNR	0.3319	0.4482	0.2466	0.7248	0.6334	0.7062
FPR	0.0164	0.0131	0.0528	0.0026	0.0020	0.0014

Table 2 Model Metrics Without Upsampling

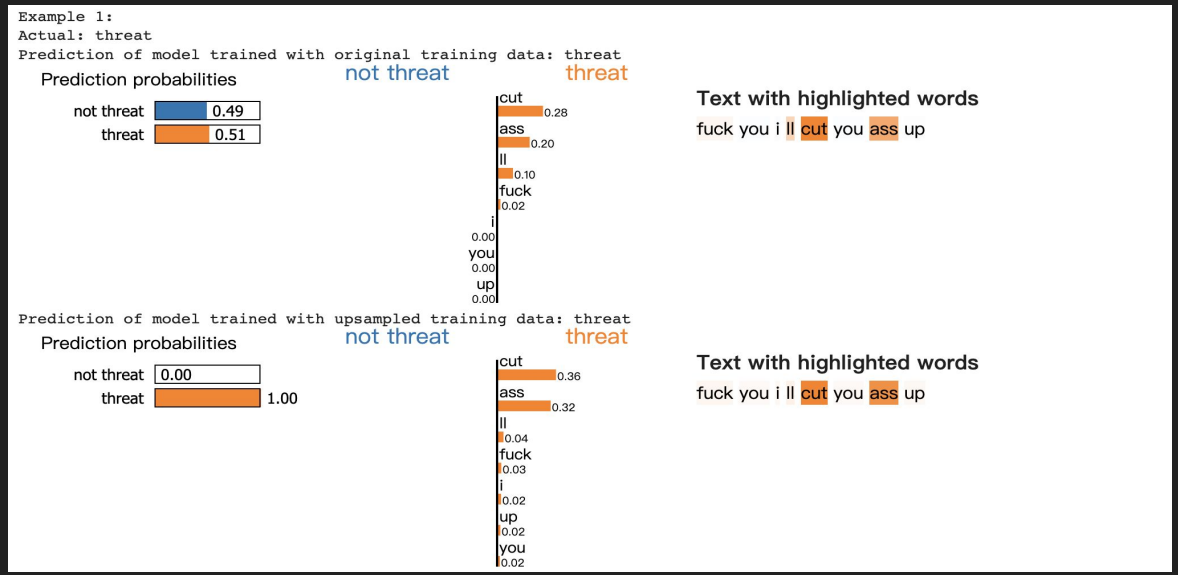
	Obscene	Insult	Toxic	Severe Toxic	Identity Hate	Threat
Accuracy	0.8734	0.8539	0.8263	0.9127	0.9070	0.9910
Recall	0.8648	0.8369	0.9095	0.6185	0.6798	0.6066
Precision	0.2957	0.2461	0.3440	0.0400	0.0780	0.2071
FNR	0.1352	0.1631	0.0905	0.3815	0.3202	0.3934
FPR	0.1261	0.1451	0.1825	0.0856	0.0905	0.0077

Table 3 Model Metrics With Upsampling

OUTCOMES

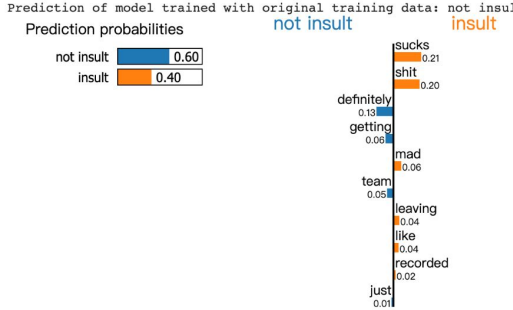
LIME

LIME is used to see how our models make the decisions. And it helps to compare the performance of the models (with and without upsampling).



Actual: insult

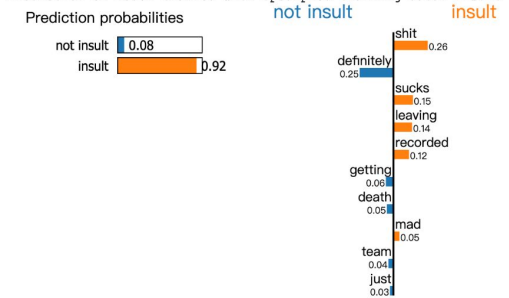
Prediction of model trained with original training data: not insult (misclassification)



Text with highlighted words

rage quitting it most definitely is not leaving just so you death doesnt get recorded or w e its getting mad and leaving anything just because your playing like shit or your team sucks

Prediction of model trained with upsampled training data: insult

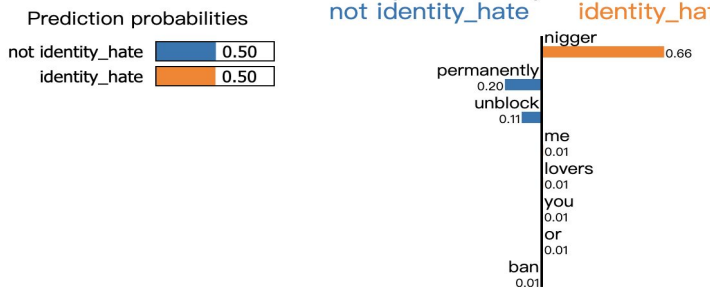


Text with highlighted words

rage quitting it most definitely is not leaving just so you death doesnt get recorded or w e its getting mad and leaving anything just because your playing like shit or your team sucks

Actual: identity_hate

Prediction of model trained with original training data: not identity_hate (misclassification)



Text with highlighted words

unblock me or ban me permanently you nigger lovers

Prediction of model trained with upsampled training data: identity_hate

Text with highlighted words

unblock me or ban me permanently you nigger lovers

SUMMARY

Problems of Data

Highly imbalanced, relatively small size

Stakeholders

Platform and users

Further Improvements

More advanced models, better text cleaning techniques, and more high-quality data